DOCUMENT RESUME

ED 405 359 TM 026 195

AUTHOR Russell, Michael; Haney, Walt
TITLE Testing Writing on Computers: Results of a Pilot Study To Compare Student Writing Test Performance via Computer or via Paper-and-Pencil.
INSTITUTION Boston Coll., Chestnut Hill, MA. Center for the Study of Testing, Evaluation, and Educational Policy.
PUB DATE 4 Oct 96
NOTE 26p.; Paper presented at the Mid-Atlantic Alliance for Computers and Writing Conference (1996).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Computer Assisted Testing; *Essay Tests; High Schools; *High School Students; Multiple Choice Tests; Performance Factors; Scoring; *Test Format; Testing Problems; *Writing Tests
IDENTIFIERS *Open Ended Questions; *Paper and Pencil Tests

ABSTRACT

The results of a small research project that studied the effect computer administration has on student performance for writing or essay tests are presented. The introduction of computer-administered tests has raised concern about the equivalence of scores generated by computer versus paper-and-pencil test versions. For this study a sample of students from the Advanced Laboratory School in Worcester, Massachusetts completed performance writing items and items from the National Assessment of Educational Progress in the traditional paper-and-pencil form or in a computerized version developed for the study. The actual number of students who participated was 42 in the computer test group and 47 in the traditional test group. The final test consisted of a combination of multiple-choice and open-ended items. To score the performance writing item, all hand-written responses were entered verbatim into the computer to prevent raters from knowing which responses were originally written by hand. Unlike much previous research on the effects of computer administration of tests, these results indicate substantial effects due to mode of administration. The size of effects was found to be larger on open-ended writing tasks than on multiple-choice tests, and the effect of mode of administration was particularly large on the extended writing task. A textual analysis of student responses to the extended writing task shows that students using the computer tended to write almost twice as much and are more apt to organize their responses into paragraphs. (Contains 6 tables and 18 references.) (SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED 405 359

**BOSTON COLLEGE**
CHESTNUT HILL, MASSACHUSETTS 02167

Center for the Study of Testing, Evaluation
and Educational Policy
323 Campion Hall
Chestnut Hill, MA 02167
Ph. 617/552-4521
Fax 617-552-8419

# Testing Writing on Computers: Results of a Pilot Study to Compare Student Writing Test Performance via Computer or via Paper-and-Pencil*

**Paper presented at the**
**Mid-Atlantic Alliance for Computers & Writing Conference**

Michael Russell and

Walt Haney

October 4, 1996

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
MICHAEL RUSSELL
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

* We wish to acknowledge our great appreciation of Carol Shilinsky, Principal, and teachers of the ALL School of Worcester, MA, who first suggested the idea of the study reported here, and in particular to Rich Tamalavich and Deena Kelly who helped arrange the equipment needed and oversaw administration of the computerized testing.

TM026195



BEST COPY AVAILABLE

## Introduction

Two of the most prominent movements in education over the last decade or so are the introduction of computers into schools and the increasing use of "authentic assessments". A key assumption of the authentic assessment movement is that instead of simply relying on multiple choice tests, assessments of students should be based on the responses students generate for open-ended "real world" tasks. "Efforts at both the national and state levels are now directed at greater use of performance assessment, constructed response questions and portfolios based on actual student work" (Barton & Coley, 1994, p. 3). At the state level, the most commonly employed kind of non-multiple-choice test has been the writing test (Barton & Coley, 1994, p. 31) in which students write their answers long-hand. At the same, many test developers have explored the use of computer administered tests, but this form of testing has been limited almost exclusively to multiple-choice tests. Relatively little attention has been paid to the use of computers to administer tests which require students to generate responses for open-ended items.

The consequences of the incongruities in these developments may be substantial. As the use of computers in schools and homes increases and students do more of their writing with word processors, at least two problems arise. First, performance tests which require students to generate responses long-hand via paper-and-pencil (which happens not just with large scale tests of writing, but also for assessments of other skills as evidenced through writing) may violate one of the key assumptions of the authentic assessment movement. For people who do most of their writing via computer, writing long-hand via paper-and-pencil is an artificial rather then real world task. Second, and more importantly, tests which require answers written long-hand with paper-and-pencil to assess students' abilities (in writing or in other subjects) may yield underestimates of the actual abilities of students who are accustomed to writing via computer.

In this paper, we present the results of a small research project which studied the effect computer administration has on student performance for writing or essay tests. Specifically, we discuss the background, design and results of the study reported here. However, before focusing on the study itself, we present a brief summary of recent developments in computerized testing and authentic assessment.

In 1968, Bert Green, Jr., predicted "the inevitable computer conquest of testing" (Green, 1970, p. 194). Since then, other observers have envisioned a future in which "calibrated measures embedded in a curriculum . . . continuously and unobtrusively estimate dynamic changes in student proficiency" (Bunderson, Inouye & Olsen, 1989, p. 387). Such visions of computerized testing, however, are far from present reality. Instead, most recent research on computerized testing has focused on computerized adaptive testing, typically employing multiple-choice tests. Perhaps the most widely publicized application of this form of testing occurred in 1993 when the Graduate Record Examination (GRE) was administered nationally in both paper/pencil and computerized adaptive forms.

Naturally, the introduction of computer administered tests has raised concern about the equivalence of scores generated by computer versus paper-and-pencil test versions. Although exceptions have been found, Bunderson, Inouye & Olsen (1989) summarize the general pattern of findings from several studies which examined the equivalence of scores acquired through computer or paper-and-pencil test forms as follows: "In general it was found more frequently that the mean scores were not equivalent than that they were equivalent; that is the scores on tests administered on paper were more often higher than on computer-administered tests." However, the authors also state that "[t]he score differences were generally quite small and of little practical significance" (p. 378). Such equating studies, however, have generally focused on the equivalence of scores obtained on standardized, multiple-choice tests. More recently, Mead & Drasgow (1993) reported on a

meta-analysis of 29 previous studies of the equivalence of computerized and paper-and-pencil cognitive ability tests (involving 159 correlations between of computerized and paper-and-pencil test results). Though they found that computerized tests were very slightly harder than paper-and-pencil tests (with an overall cross-mode effect size of -.04), they concluded that their results "provide strong support for the conclusion that there is no medium effect for carefully constructed power tests. Moreover, no effect was found for adaptivity. On the other hand, a substantial medium effect was found for speeded tests" (Mead & Drasgow, 1993, p. 457)

Yet, as previously noted, standardized multiple-choice tests, which have been the object of comparison in previous research on computerized versus paper-and-pencil testing, have been criticized by proponents of authentic assessment. Among the characteristics which lend authenticity to an assessment instrument, Darling-Hammond, Ancess & Falk (1995) argue that the tasks be "connected to students' lives and to their learning experiences..." and that they provide insight into "students' abilities to perform 'real world' tasks" (p.4-5). Unlike standardized tests, which typically are viewed as external instruments that measure a fraction of what students have learned, authentic assessments are intended to be closely linked with daily classroom activity so that they seamlessly "support and transform the process of teaching and learning" (Darling-Hammond, Ancess & Falk, 1995, p. 4; Cohen, 1990).

In response to this move towards authentic assessment, many developers of nationally administered standardized tests have attempted to embellish their instruments by including open-ended items. These changes, however, have occurred during a period when both the real-world and the school-world have experienced a rapid increase in the use of computers.

The National Center for Education Statistics estimates that the percentage of students in grades 1 to 8 using computers in school has increased from 31.5 in 1984, to 52.3 in

1989 and to 68.9 in 1993 (Snyder & Hoffman, 1990; 1994). In the workplace, the percentage of employees using computers has risen from 36.0 in 1989 to 45.8 in 1993. Moreover, during this time period, writing has been the predominant task adult workers perform on a computer (Snyder & Hoffman, 1993; 1995). Given these trends, tests which require students to generate responses to open-ended items via paper-and-pencil may decrease the test's "authenticity" in two ways: 1. Assessments are not aligned with students' learning experiences; and 2. Assessments are not representative of 'real-world' tasks. As the remainder of this paper suggests, these shortcomings may be leading to underestimates of students' writing abilities.

## Background to this Study

In the fall of 1993, the Advanced Learning Laboratory School (ALL School) of Worcester, Massachusetts decided to adopt the Co-NECT school design (or Cooperative Networked Educational Community for Tomorrow). Developed by Bolt Beranek and Newman Inc., a Boston-based communications technology firm, Co-NECT is one of nine models for innovative schooling funded by the New American Schools Development Corporation. Working with Bolt, Beranek and Newman (BBN), the ALL School restructured all aspects of its educational environment. Among other reforms, the traditional middle school grade structure (that is separately organized grade 6, 7 and 8 classes) was replaced with blocks which combined students who otherwise would be divided into grades 6, 7 and 8 into a single cluster. In place of traditional subject-based classes (such as English Class, Math Class, Social Studies, etc.), all subjects were integrated and were taught through project-based activities. To support this cooperative learning structure, several networked computers were placed in each classroom, allowing students to perform research via the Internet and CD-ROM titles, to write reports, papers and journals, and to create computer based presentations using several software applications.

To help evaluate the effects the restructuring at the ALL School has on its students as a whole, the Center for the Study of Testing, Evaluation and Educational Policy at Boston College helped teachers gather baseline data in the Fall of 1993 with plans to perform follow-up assessments in the spring of 1994 and each spring thereafter. To acquire a broad picture of students' strengths and weaknesses, the forms of tests included in the baseline assessment ranged from multiple choice tests to short and long answer open-ended assessments to hands-on performance assessments covering a wide range of reading, writing, science and math skills. To acquire insight into how cooperative projects affected the development of group skills, some of the performance assessments required students to work together to solve a problem and/or answer specific questions. Finally, to evaluate how the Co-NECT Model, as implemented in the ALL School, affected students' feelings about their school, a student survey was administered to a sample of students.

In the spring of 1994, the same set of assessments was re-administered to a different representative sample of students. While a full discussion of the results is beyond the scope of this paper, many of the resulting patterns of change were as expected. For example, performance items which required students to work cooperatively generally showed more improvement than items which required students to work independently. On items that required students to work independently, improvement was generally stronger on open-ended items than on multiple-choice items. But there was one notable exception: open-ended assessments of writing skills suggested that writing skills had declined.

Although teachers believed that the Co-NECT Model enhanced opportunities for students to practice writing, performance on both short answer and long answer writing items showed substantial decreases. For example, on a short answer item which asked students to write a recipe for peace, the percentage of students who responded satisfactorily decreased from 69% to 51%. On a long answer item which asked students to imagine a superhero, describe his/her powers, and write a passage in which the superhero uses

his/her powers, the percentage of satisfactory responses dropped from 71% to 41%. On another long answer item that asked students to write a story about a special activity done with their friends or family, student performance dropped from 56% to 43%. And on a performance writing item which first asked students to discuss what they saw in a mural with their peers and then asked them to independently write a passage that described an element in the mural and explain why they selected it, the percentage of satisfactory responses decreased from 62% to 47%.

Since writing was a skill the school had selected as a focus area for the 1993-94 school year, teachers were surprised and troubled by the apparent decrease in writing performance. During a feedback session on results in June 1994, teachers and administrators discussed at length the various writing activities they had undertaken over the past year. Based on these conversations, it was evident that students were regularly presented with opportunities to practice their writing skills. But consistent throughout many of the comments was that teachers in the ALL School were increasingly encouraging students to use computers and word processing tools in their writing. As several computers were present in all classrooms, as well as in the library, teachers believed that students had become accustomed to writing on the computer. When one teacher suggested that the decrease in writing scores might be due to the fact that all writing items in spring 1994 were administered on paper and required students to write their responses by hand, the theory was quickly supported by many teachers. With a follow-up assessment scheduled to occur a year later, several teachers asked if it would was possible for students to perform the writing items on computer.

After careful consideration, it was decided that a sub-sample of students in spring 1995 would perform the performance writing item and items from the National Assessment of Education Progress or NAEP (which included mostly multiple-choice items, but also several short answer items) on the computer. But, to preserve comparisons with results

from 1993-94, the majority of the student population would perform these assessments as they had in that year -- via the traditional pencil-and-paper medium.

## Study Design and Test Instruments

To study the effect the form of administration has on student performance, that is writing items on computer versus by hand on paper, two groups of 50 students were randomly selected from the ALL School Advanced Cluster (grades 6, 7 and 8). These students performed three assessments:

1. An open-ended assessment comprising 14 items, which included two writing items, five science items, five math items and two reading items.

2. A sub-test of the National Assessment of Educational Progress (NAEP) which was divided into three sections which included 15 language arts items, 23 science items and 18 math items. The majority of NAEP items were multiple-choice. However, 2 language arts items, 3 science items and 1 math item were open-ended and required students to write a brief response to each items prompt.

3. A performance writing assessment which required an extended written response.

The performance writing assessment consisted of a picture of a mural and two questions. Students formed small groups of 2 or 3 to discuss the mural. After 5 to 10 minutes, students returned to their seats and responded to one of two prompts:

1. Now, it is your turn to pick one thing you found in the mural. Pick one thing that is familiar to you, that you can recognize from your daily life or that is part of your culture. Describe it in detail and explain why you chose it.

2. Artists usually try to tell us something through their paintings and drawings. They may want to tell us about their lives, their culture or their feelings about what is happening in the neighborhood, community or world. What do you think the artists who made this mural want to tell us? What is this mural's message?

For comparative purposes, all students performed the open-ended assessment via the traditional pencil-and-paper medium. The experimental group of students performed the NAEP and the performance writing assessments on computer, while the control group performed the NAEP and the performance writing assessments by hand on paper.

Due to absences, the actual number of students who participated in this study was as follows:

Computer Test Group - 42

Paper-and-pencil Test Group - 47

## Converting Paper Tests to Computer

Before the tests could be administered on computer, the paper versions were converted to a computerized format. Several studies suggest that slight changes in the appearance of an item can affect performance on that item. Something as simple as changing the font in which a question is written, the order items are presented, or the order of response options can affect performance on that item (Beaton & Zwick, 1990; Cizek, 1991). Other studies have shown that people become more fatigued when reading text on a computer screen than when they read the same text on paper (Mourant, Lakshmanan & Chantadisai, 1981). One study (Haas & Hayes, 1986) found that when dealing with passages that covered more than one page, computer administration yielded lower scores than paper-and-pencil administration, apparently due to the difficulty of reading extended

text on screen. Clearly, by converting items from paper to computer, the appearance of items is altered.

To minimize such effects, each page of the paper version of the NAEP items and performance writing item was replicated on the computer screen as precisely as possible. To the extent possible, the layout of text and graphics on the computer version matched the paper version, including the number of items on a page, the arrangement of response options, and the positioning of footers, headers and directions. Despite these efforts, not every screen matched every page. Since the computer screen contained less vertical space, it was not always possible to fit the same number of questions on the screen as appeared on the page. In addition, to allow the test taker to move between screens (e.g., to go on to the next screen, back to a previous screen, or to flip to a passage or image to which an item referred), each screen of the computer versions contained navigation buttons along its bottom edge. Finally, to decrease the impact of screen fatigue, a larger font was used on the computer version than on the paper version.

To create a computerized version of the NAEP and performance writing tests, the following steps were taken:

1. An appropriate authoring tool was selected. To fully integrate the several graphics used in the multiple-choice items and the full-color photograph of a mural used in the performance writing item, as well as to track students responses, Macromedia Director was used.

2. All graphics and the photograph of the mural were scanned. Adobe Photoshop was used to retouch the images.

3. A data file was created to store student inputs, including their name, ID number, school name, birthdate, gender, date of administration and responses to each item.

4. A prototype of each test was created, integrating the graphics, text and database into a seamless application. As described earlier, navigational buttons were placed along the lower edge of the screen. In addition, a "cover" page was created in which students entered biographical information.

5. The prototype was tested on several adults and students to assure that all navigational buttons functioned properly, that data was stored accurately, and that items and graphics were easy to read.

6. Lastly, the prototype was revised as needed and the final versions of the computer tests were installed on twenty-four computers in the ALL School.

As described above, the addition of navigational buttons along the lower edge of the computer screen was the most noticeable difference between the computer and paper versions of the tests. To allow students to review their work and make changes as desired, a "Next Page" and "Previous Page" button appeared on all pages (or screens) of the computer tests (except the first and last page). To allow students to review their work, student responses were not recorded until the student reached the last page of the assessment and clicked a button labeled "I'm Finished." When the "I'm Finished" button was clicked, the student's biographical information and responses to each item were recorded in a data file before the program terminated. For all multiple-choice items, students clicked the option they felt best answered the question posed. For both short- and long-answer questions, examinees used a keyboard to type their answers into text boxes which appeared on their screen. Though they could edit using the keyboard and mouse, examinees did not have access to word processing tools such as spell-checking.

## Scoring

A combination of multiple choice and open-ended items were performed by both groups of students. Multiple-choice NAEP items were scored as either correct or incorrect

based upon the answer key accompanying the NAEP items. To prevent rater bias based on the mode of response, all short-answer NAEP responses were entered verbatim into the computer. Responses of students who had taken the NAEP questions on computer and via paper-and pencil were then randomly intermixed. Applying the rating rubrics designed by NAEP, two raters independently scored each set of six short answer items for each student. After each rater finished scoring all items for all students, their ratings (which ranged from 1 - 5) were converted to a dichotomous value: 1 or 0; to denote whether student responses were adequate or inadequate. The two raters' converted scores were then compared. Where discrepancies occurred, the raters re-evaluated responses and reached consensus on a score.

To score the performance writing item, all hand written responses were entered verbatim into the computer -- again so as to prevent raters from knowing which responses were originally written by hand. The hand written and computer responses were randomly intermixed. Three independent raters then scored each written response, using the following four-point scoring rubric:

1 - Too brief to evaluate: student did not make an attempt, indicates that student did not know how to begin, or that the student could not approach the problem in an appropriate manner.

2 - Inadequate Response: student made an attempt but the response was incorrect, reflected a misconception and/or was poorly communicated.

3 - Adequate Response: response is correct and communicated satisfactorily, but lacks clarity, elaboration and supporting evidence.

4 - Excellent Response: response is correct, communicated clearly and contains evidence which supports his/her response.

Initial analyses of the three raters' ratings showed that there was only a modest level of inter-rater reliability among the three (inter-rater correlations ranged from 0.43 to 0.63,

across the total of 89 performance writing responses). Although these correlations were lower than expected, research on the assessment of writing has shown that rating of writing samples, even among trained raters, tends to be only modestly reliable (Dunbar, Koretz, & Hoover, 1991). Indeed, that is why we planned to have more than one rater evaluate each student response to the performance writing task. Hence for the purpose of the study reported here we created composite performance rating scores by averaging the three ratings of each student's response (which we call PWAvg).

Since the open-ended assessment was performed by hand by all students, student responses were not entered into the computer. A single rater, who did not know which students had performed other assessments on the computer, scored all responses using a 4 point scale. Although each of the 14 items had its own specific scoring criteria, the general meaning of each score was the same across all 14 open-ended items, as well as the performance writing item. The raw scores were then collapsed into a 0, 1 scale, with original scores of 1 or 2 representing a 0, or inadequate response, and original scores of 3 or 4 representing a 1, or adequate response. For the purpose of the study reported here, total open-ended response scores were calculated by summing across all 14 OE items.

## Results

In presenting results from this study, we discuss: 1) assessment results overall; 2) comparative results from the two groups who took assessments via computer and via paper-and-pencil; 3) results of regression analyses; and 4) a comparison of the speededness of the two modes.

### Overall Results

Table 1 presents a summary of overall results, that is, combined results for all students who took any of the three assessments in Spring 1995.

Table 1: Summary Statistics for All Assessments

| | Scale Range | n | Mean | SD |
|---|---|---|---|---|
| **OE** | 0 - 14 | 114 | 7.43 | 3.10 |
| **NAEP Lang Arts** | 0 - 15 | 120 | 7.23 | 5.40 |
| **NAEP Science** | 0 - 23 | 120 | 7.16 | 5.70 |
| **NAEP Math** | 0 - 18 | 120 | 4.62 | 3.98 |
| **Perf Writing Avg** | 1 - 4 | 89 | 2.53 | 0.62 |

These data indicate that the assessments were relatively challenging for the students who performed them. Mean scores were in the range of 50% correct for the OE and NAEP Language Arts tests, but considerably below 50% correct for the NAEP science and NAEP math subtests. In this regard, it should be noted that all of these assessments were originally designed to be administered to eighth graders, but in the study reported here were administered to 6th, 7th and 8th grade level students who in the ALL school are intermixed in the same clusters.

Table 2 presents Spearman rank order intercorrelations of all assessments, again across both experimental groups. The OE results correlated most highly with the PW Avg results, likely reflecting the fact that both of these assessments were open-ended requiring students to produce rather than select an answer. The three NAEP item subtests showed fairly high intercorrelations (0.80-0.84) which might be expected for multiple-choice tests in the different subject areas (despite the fact that none of the NAEP subtests contained as many as two dozen items). The PW Avg results showed only modest correlations with the NAEP subtests, but this is not unexpected given that the PW scores are based on an extended writing task, while the NAEP results are based largely on multiple choice items.

Table 2: Intercorrelations of Assessment Results

| | OE | NAEP Lang Arts | NAEP Science | NAEP Math | Perf Writing |
|---|---|---|---|---|---|
| **OE** | 1.000 | | | | |
| **NAEP Lang Arts** | 0.172* | 1.000 | | | |
| **NAEP Science** | 0.221** | 0.839*** | 1.000 | | |
| **NAEP Math** | 0.145 | 0.804*** | 0.805*** | 1.000 | |
| **Perf Writing** | 0.475*** | 0.441*** | 0.464*** | 0.438*** | 1.000 |

* $p < .05$, ** $p < .01$, *** $p < .001$

### Computer versus Paper-and-Pencil Results

Table 3 presents results separately for the two experimental groups, namely the group which took assessments on paper and the one that took them on computer. The table also shows results of t-tests (for independent samples, assuming equal variances for the two samples and hence using a pooled variance estimate). As an aid to interpretation, the table also shows the effect of computer administration in terms of effect size, that is the mean of

the experimental group minus the mean of the control group divided by the standard deviation of the control group.

Results indicate that for NAEP subtests and for the PW Avg, the effects of computer administration were highly significant and the magnitudes of effects were quite substantial, ranging from 0.54 to 0.96. The effect of computer administration was largest on the performance writing task and smallest on the NAEP math test (on which all but one item was multiple-choice).

The only test which did not show a statistically significant difference between the two groups was the OE test. But this was of course expected since this was the one test that was administered in the same form (paper-and-pencil) to the two groups. Note however that the size of the difference in OE scores between the two groups was surprisingly large, given that the two groups had been randomly selected. The absence of six students who participated in the main experiment but did not perform the open ended test may partially explain this difference. Nevertheless to explore the possibility that group differences may partially account for apparent mode of administration effects (and also, of course to estimate effects more precisely), regression analyses were conducted.

Table 3: Summary Results by Group

| | Paper/Pencil | | | Computer | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | n | Mean | SD | n | Mean | SD | Effect size | t (df) | Sig |
| **OE** | 68 | 6.88 | 3.27 | 46 | 8.24 | 2.66 | 0.42 | 2.34 (112) | 0.02 |
| **NAEP Lang Arts** | 72 | 5.68 | 5.43 | 48 | 9.54 | 4.49 | 0.71 | 4.08 (118) | <0.0001 |
| **NAEP Science** | 72 | 5.40 | 5.39 | 48 | 9.79 | 5.15 | 0.81 | 4.45 (118) | <0.0001 |
| **NAEP Math** | 72 | 3.78 | 3.90 | 48 | 5.88 | 3.80 | 0.54 | 2.91 (118) | <0.0043 |
| **Perf Writing** | 47 | 2.28 | 0.55 | 42 | 2.81 | 0.58 | 0.96 | 4.40 (87) | <0.0001 |

## Regression Analyses

As a further step in examining the effects of mode of administration, a series of regression analyses were conducted using the OE scores as a covariate and then introducing a dummy variable (0= paper/pencil group; 1= computer administration group) to estimate the effects of mode of administration on the NAEP Language Arts, Science and Math subtests and on the PW Avg scores. Results of these regression analyses are shown in Table 4.

Table 4: Results of Regression Analyses

| Dependent Var | | Coefficient | s.e. | t-ratio | Prob |
|---|---|---|---|---|---|
| **NAEP Lang Arts** | Constant | 4.90 | 1.23 | 4.00 | 0.0001 |
| | OE | 0.11 | 0.15 | 0.68 | 0.4963 |
| | Group (1=comp) | 4.18 | 0.97 | 4.29 | <0.0001 |
| **NAEP Science** | Constant | 3.41 | 1.24 | 2.74 | 0.0071 |
| | OE | 0.27 | 0.16 | 1.70 | 0.0923 |
| | Group (1=comp) | 4.60 | 0.99 | 4.65 | <0.0001 |
| **NAEP Math** | Constant | 2.45 | 0.93 | 2.64 | 0.0095 |
| | OE | 0.20 | 0.12 | 1.66 | 0.0994 |
| | Group (1=comp) | 2.07 | 0.74 | 2.81 | 0.0059 |
| **Perf Writing** | Constant | 1.61 | 0.17 | 9.71 | <0.0001 |
| | OE | 0.09 | 0.02 | 4.77 | <0.0001 |
| | Group (1=comp) | 0.43 | 0.11 | 3.89 | 0.0002 |

These results essentially confirm the findings shown in Table 3, namely that even after controlling for OE scores, the effect of mode of administration was highly significant on all three NAEP subtests and also on the PW Avg.

## Speededness

As noted above, scores on the three NAEP subtests (Language Arts, Science, and Math) differed significantly between the experimental and control groups. Most items on the NAEP sub-tests, however, were multiple choice. Although we expected the use of a computer might lead to higher performance on open-ended items, in light of previous research on the equivalence of (Bunderson, Inouye & Olsen, 1989; Mead & Drasgow 1993), we did not expect the form of test to affect performance on multiple choice items.

To investigate possible causes for differences in performance on multiple choice items, in light of the findings of Mead & Drasgow (1993), we examined the effect the form of the test had on the speededness of the test. As Donlon (1984) reports, the College Board Admissions Testing Program has long had the policy of regarding a test as essentially unspeeded "if at least 80 percent of the group reach the last item and if everyone reaches at least 75 percent of the items" (p. 36). Thus to investigate speededness, we examined the proportions of students who reached 75%, 90% and 100% of the items on the NAEP subtests.

Table 5 presents the results of our analysis of the speededness of the two forms. The table also shows the results of chi-square goodness of fit tests along with the corresponding probabilities. Results indicate that there was a significant difference in the speededness of the two forms for the Language Arts and Science sub-tests, but not for the Math sub-test.

While 89% of the control group reached the first 75% of the Language Arts items and only 82% of the control group reached the first 90% of the items, all of the students in the experimental group reached at least the first 90% of the items. And when all items are considered, only 52% of the control group reached the last item compared with 91% of the

students who performed the test on computer. These differences are significant at the 95% confidence level.

Table 5: Results of Analysis of Speededness

| | Percent of Students Reaching | | | |
|---|---|---|---|---|
| | Paper Version | Computer Version | Chi-Square | p |
| **Language Arts** | | | | |
| **reaching 75%** | 89% | 100% | 5.3 | <.05 |
| **reaching 90%** | 82% | 100% | 8.72 | <.01 |
| **reaching 100%** | 52% | 91% | 16.15 | <.001 |
| **Science** | | | | |
| **reaching 75%** | 93% | 98% | 1.05 | NS |
| **reaching 90%** | 73% | 91% | 4.88 | <.05 |
| **reaching 100%** | 50% | 82% | 9.91 | <.01 |
| **Math** | | | | |
| **reaching 75%** | 95% | 93% | .22 | NS |
| **reaching 90%** | 93% | 86% | 1.06 | NS |
| **reaching 100%** | 86% | 77% | 2.0 | NS |

On the Science sub-test, there was no significant difference for the first 75% of the items. However, only 73% of the control group reached 90% of the items and only 50% of these students reached all of the items, as compared to 91% and 82% of the experimental group, respectively. Again, these differences are significant at the 95% confidence level.

On the Math sub-test, however, significant differences in speededness were not detected. It seems, then, that a difference in the speededness of the two forms does not account for the observed difference between the mean performance on the Math sub-test for the two groups. However, since items that were not reached were considered incorrect, it is likely that the control group's mean scores on both the Language Arts and Science sub-tests were adversely affected by the observed differences in speededness. Similarly, since four of the five open-ended items on these two tests appeared among the last 10% of items,

the differences in performance on these open-ended questions is also partially explained by differences in speededness. However, as item 8 on the Language Arts subtest and item 17 on the Math subtest suggest, speededness only accounts for a portion of the difference in performance on open-ended NAEP items. Table 6 shows that for item 8, which was reached by all but one student, only 35% of the control responded adequately while 75% of the computer group responded adequately. Similarly, Table 7 indicates that a significant difference in performance also occurred on item 17 of the Math subtest, for which speededness did not differ significantly. Whereas only 23% of the control group responded adequately, 57% of the experiment group responded adequately on this item.

Table 6: Results on Language Arts Subtest Item 8

| | % Correct | t statistic | p |
|---|---|---|---|
| **Paper Version** | 35% | | |
| **Computer Version** | 75% | 8.83 | <.001 |

Table 7: Results on Math Subtest Item 17

| | % Correct | t statistic | p |
|---|---|---|---|
| **Paper Version** | 23% | | |
| **Computer Version** | 57% | 7.2 | <.001 |

In considering differences in speededness, it should be noted that there were four separate administrations of the NAEP test, each administered by a different person. Because this was the first time the computer version had been used, it is possible that the test administrators for the computer version were more lenient in allotting time to work on the test. In one computer administration, it was noted that although the test was designed to take 45 minutes, the students remained in the room for just over an hour. While this additional time might be accounted for by a tardy start or by allowing students to rest between sections, it may have contributed to the perceived differences in the speededness of the two versions.

## Discussion

The study described here is best characterized as a small exploratory inquiry. Motivated by a question as to whether or not student results on an extended writing task might be better if they were allowed to write on computer rather than by hand, the study was aimed at estimating the effects of mode of administration on test results for two kinds of assessments, namely the largely multiple-choice NAEP subtests and the extended writing task previously described. Unlike much previous research on the effects of computer administration of tests, which has largely focused on multiple-choice tests and which generally found no or small differences due to mode of administration, our results indicate substantial effects due to mode of administration. The size of the effects was found to be larger on open-ended writing tasks than on multiple-choice tests. The effect of mode of administration was particularly large on the extended writing task - where we found an effect size in the range of 0.80-0.95. Effect sizes of this magnitude are unusually large. An effect size of 0.80, for example, implies that the score for the average student in the experimental group exceeds that of 79 percent of the students in the control group.

As a means of inquiring further into the source of this large effect, we conducted a textual analysis of student responses to the extended writing task. Specifically we calculated the average number of words and paragraphs contained in the responses of both groups. As Table 8 below indicates, those students who performed the assessment on the computer tended to write almost twice as much and were more apt to organize their responses into more paragraphs.

Table 8: Characters, Words and Paragraphs on Performance Writing Task by Mode of Administration

| | Characters | Words | Paragraphs |
|---|---|---|---|
| **Computer** | | | |
| **Mean** | 1007.4 | 201.9 | 2.628 |
| **Std** | 535.58 | 108.5 | 2.236 |
| **n** | 43 | 43 | 43 |
| **Paper** | | | |
| **Mean** | 577.5 | 110.0 | 1.438 |
| **Std** | 274.81 | 52.11 | 1.05 |
| **n** | 48 | 48 | 48 |
| **observed t with pooled variance** | 4.89 | 5.23 | 3.31 |

In some regards, this pattern is consistent with the findings of Daiute (1985) and Morocco (1987), who have shown that teaching writing with word processors tends to lead students to write more and to revise more than when they write by hand. What is clear, however, is the apparent differences in the quality of these responses as suggested by the higher scores generally awarded to responses generated on the computer (as rated by scorers who did not know which responses had originally been done on computer and by hand).

Although our small study had several weaknesses — only one extended writing task was used, no other data beyond the OE test results were used as covariates in regression analyses, and information on individual students' amount of experience working on computers was not collected — further research into this topic clearly is warranted.

Increasingly, schools are encouraging students to use computers in their writing. As a result, it is likely that an increasing number of students are growing accustomed to writing on computers. Nevertheless, large scale assessments of writing, at state, national and even international levels, are attempting to estimate students' writing skills by having them write by hand on paper. Our results, if generalizable, suggest that for students

accustomed to writing on computer, such estimates of student writing abilities based on hand writing may be substantial underestimates of their abilities to write when using a computer.

## References

Barton, P. E. & Coley R. J. (1994) *Testing in America's schools*. Princeton, NJ Educational Testing Service Policy Information Center.

Beaton, A. E. & Zwick, R. (1990). *The Effect of Changes in the National Assessment: Disentangling the NAEP 1985-86 Reading Anomaly*. Princeton, NJ: Educational Testing Service, ETS.

Bunderson, C. V., Inouye, D. K. & Olsen, J. B. (1989). The four generations of computerized educational measurement. In Linn, R. L., *Educational Measurement* (3rd Ed). Washington, D.C.: American Council on Education, pp. 367-407.

Cizek, G. J. (1991). The Effect of Altering the Position of Options in a multiple-choice Examination. Paper presented at NCME, April 1991. (ERIC)

Cohen, D. (1990). Reshaping the Standards Agenda: From an Australian's Perspective of Curriculum and Assessment. In P. Broadfoot, R. Murphy & H. Torrance (Eds.), *Changing Educational Assessment: International Perspectives and Trends*. London: Routledge.

Daiute, C. (1985). *Writing and computers*. Reading, MA: Addison-Wesley Publishing Co.

Darling-Hammond, L., Acness, J. & Falk, B. (1995). *Authentic Assessment in Action*. New York, NY: Teachers College Press.

Donlon, T. (Ed) (1984).*The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests*. NY: College Entrance Examination Board.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality Control in the Development and Use of Performance Assessments. *Applied Measurement in Education*, 4(4) 289-303.

Green, B. F., Jr. (1970). Comments on tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing and guidance*. New York: Harper and Row.

Haas, C. & Hayes, J. R. (1986). What Did I Just Say? Reading Problems in Writing with the Machine. Research in the Teaching of English, 20:1 22-35.

Mead, A. D & Drasgow, F. (1993) Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114:3 449-458.

Morocco, C. C. & Neuman, S. B. (1986). Word processors and the acquisition of writing strategies. *Journal of Learning Disabilities* 19(4) 243-248.

Mourant, R. R, Lakshmanan, R. & Chantadisai, R. (1981). Visual Fatigue and Cathode Ray Rube Display Terminals. *Human Factors*, 23(5), 529-540.

Snyder, T. D. & Hoffman, C. M. (1990). *Digest of Education Statistics*. Washington, DC: U. S. Department of Education.

Snyder, T. D. & Hoffman, C. M. (1993). *Digest of Education Statistics.* Washington, DC: U. S. Department of Education.

Snyder, T. D. & Hoffman, C. M. (1994). *Digest of Education Statistics.* Washington, DC: U. S. Department of Education.

Snyder, T. D. & Hoffman, C. M. (1995). *Digest of Education Statistics.* Washington, DC: U. S. Department of Education.